The Effect of High-speed Railways on Knowledge Transfer: Evidence from Japanese Patent Citations [*]

Ryuichi Tamura

Faculty of Integrated Media, Wakkanai Hokusei Gakuen University

Abstract

This paper tries to identify the causal relationships between the high-speed railway services (the Shinkansen lines of Japan) and knowledge transfer as evidenced by patent citations. Using the opening of the Hokuriku Shinkansen line in 1997 as an exogenous variation, we designed a natural experiment to assess the changes in the frequencies and geographical extent of knowledge transfer to the area where the Shinkansen runs. We apply the difference in difference methodology to examine the effect of the operation, and find that the Shinkansen enhances the knowledge transfer around the area, and expands the geographical extent in which the knowledge transfer takes place.

Keywords: knowledge transfer, patent citations, knowledge spillovers, high-speed railways, natural experiment

JEL Classification: D62, C99, L91

I. Introduction

How does new knowledge spread? Arrow (1962) states that knowledge and information spread by means of moves of human capital such as researchers and engineers. In this study, we refer to this process as *knowledge transfer*. Knowledge transfers can be broken down into the following three components: senders who invent new knowledge, receivers who exploit it, and mechanisms or mediums with which receivers acquire new knowledge. In this study, we focus on a mechanism that enhances knowledge spillovers, because as emphasized by both theoretical and empirical literature in economics, the spillover effect of knowledge is essential for economic growth.

In this study, we study the mechanisms by which knowledge is transferred as knowledge spillovers. Knowledge is usually treated as public good, but, as many empirical studies show, its geography is confined. Assuming any form of interaction between the sender and receiver, the existence of a social infrastructure that reduces the cost of knowledge transfer can play a

[*] This study is financially supported by the JST/RISTEX project. The topic of the paper is derived by the discussion with Makoto Nirei and Mark Schankerman in 2013. However, all errors remain the author's own. The author also thanks the audience at the Spatial Economics Workshop at the University of Tokyo, and the ARISH workshop at Nihon University for helpful comments.

role to expand the geographical extent. We consider the Shinkansen service of Japan, because the Shinkansen lines can promote the move of people by reducing (generalized) transport costs. For example, the Hokuriku Shinkansen line reduces the time to reach the Tokyo area from 2 hours 30 minutes to 1 hour and 19 minutes. Therefore, the opening of the Shinkansen, makes knowledge transfer from Tokyo easier. In this study, we use a patent application as an indicator of innovation and patent citations as trails of knowledge transfer, and empirically examine whether the Shinkansen line works as a transfer mechanism between senders and receivers.

Empirical study capturing knowledge spillovers in knowledge transfer begins in the context of the measurement of the spillover effect in Research and Development (R&D) investment using the "knowledge production function" approach formulated by Griliches (1979). Jaffe (1986, 1989) uses this approach to find positive correlations between the R&D investment by private firms and universities, and the productivity of private firms having similar industrial technologies in the neighboring area, suggesting that the effect of R&D spending of an organization can spill over nearby firms. Acs, et al. (1994) further applies the analysis by firm size, and found that the spillover effect is more significant for small and medium sized companies than large companies. These studies aim to study the relationships among economic variables aggregated at national or regional levels, and no information can be obtained for who is a sender, who are receivers, and how knowledge flows between them. To track knowledge flows at a micro level, Jaffe, et al. (1993) propose the use of patent citation records. They capture knowledge spillovers from inventors of cited patents to those of citing patents, which can be considered as the flow of new knowledge documented as a patent document, thereby we acquire the details of knowledge transfer in which senders and receivers are known to us from inventor name records.

Since the early 1990s, patent data administrated by the United States Patent and Trademark Office has become widely available to researchers. Patent records include not only the technology and time profile of new knowledge but also detailed information on organizations (applicants) and inventors such as their names and geographical locations. Therefore, since Jaffe, et al. (1993), numerous empirical studies in Economics, Management Science, and Information Science draw on patent records to investigate various forms of knowledge transfer or sharing. There are three strands of the empirical literature. The first literature focuses on the move of researchers who embody scientific new knowledge leading to innovations. Here, knowledge is spilt over collaborators in the joint research project by way of the inventor's move between organizations (Nakajima, et. al (2010), Breschi and Lissoni (2009), Agrawal, et. al (2006), and Saito and Yamauchi (2015) using Japanese patent records). The second focuses on the social network in which innovative researchers are embedded such as past collaborative relationships, and cultural/social connections including race and nationality (Saxenian (1999), Singh (2005), Agrawal, et. al (2008)). The third is the study on the measurement of the geographical localization of knowledge spillovers originated by Jaffe, et al. (1993) (Thompson and Fox-Kean (2005), Murata, et al. (2014)).

We learn from the third literature that the locational proximity to the source of information matters in order to receive spillovers in knowledge transfer. In this study, we raise the following two questions: Is the proximity necessary to receive knowledge spillovers? Prior to this study, the author analyzed the geographical extent of knowledge spillovers using Japanese patent citations, and estimated it to be about 60 kilometers in Japan. However, if there is a way to reduce the cost to access a knowledge source born in a distant area, researchers or engineers who are located in areas more than 60 kilometers away would be able to access the content of knowledge. Second, for them, how can be a specific channel of knowledge transfer? Empirical studies using patent citations identify a sender and receivers from the patent bibliographic information of citing and cited patents, the mechanism of which has been treated as a "black box". We posit that high speed railways in Japan is included in the black box. To be specific, we focus on the Shinkansen lines of Japan, and investigate if they serve as carriers in knowledge transfer.

To wrap up, the main hypothesis of this study is that Shinkansen reduces the (generalized) travel cost to access external knowledge in a distant area, thereby promotes inventors' moves and knowledge transfer. This hypothesis then implies that a Shinkansen line enables researchers in the area to search or access wider areas than those in the region without Shinkansen services.

I-1. Methodology

To empirically analyze the effect of a Shinkansen line on activities to source external knowledge in an area, this study makes use of events when a Shinkansen line begins operations in an area, and tries to identify the causal effect of the opening. Based on the question described above, we set the following empirical hypotheses regarding the causal relationships: First, the frequencies of knowledge transfer increase (H1). Second, the geographical extent of knowledge transfer in the area expands further (H2). Third, the direction of knowledge transfer changes from within the area to outside the area (H3). To test these hypotheses in an empirical framework, we construct a natural experiment framework which makes use of the opening of the Hokuriku (Nagano) Shinkansen in October 1997 as the exogenous event. Because the Shinkansen line was built in time for the 1998 Winter Olympics in Nagano, the assignment in the access (more specifically, the road access) to the newly built Shinkansen stations for firms and researchers in the region can be considered as independent of their innovative activities. In the area where the Hokuriku Shinkansen was to run, firms and researchers engaged in innovative activities that differed to the extent that they exploited external knowledge in distant areas: Some only use knowledge born in the area. Others frequently acquire knowledge transferred by their parent organizations located in distant innovative clusters such as Tokyo and Osaka areas where many firms actively engage in patenting activities. Such idiosyncratic differences in the geographical extent of knowledge sourcing activities among firms in the area are offset at an aggregate level by the random assignment of the route to the Hokuriku Shinkansen stations (especially for those in the area where the Shinkansen is to run).

Therefore, if there is any difference in the degree and extent of knowledge transfer before and after the opening of Hokuriku Shinkansen, it can be considered as the average treatment effect of the Shinkansen. It is also noted that the cost of knowledge transfer is affected by improvements in various forms of social capital such as road networks and the internet technologies. Assuming these improvements are made nationwide in Japan, we control for this effect by capturing the differences observed in areas that are far away from any Shinkansen station, before and after the Hokuriku Shinkansen is built. The effect, caused by factors other than the opening of the Shinkansen, is taken as a nationwide temporal change that the area along the Shinkansen line would have experienced even if the line was not to run. Based on these considerations, we set the knowledge transfers in which receivers were located around the Hokuriku Shinkansen stations as the treatment group, and those in which both a sender and receivers were located far away from any Shinkansen station as the control group. Then, the average treatment effect of the opening of Hokuriku Shinkansen can be measured by the differences between the temporal differences in the treatment group and the temporal differences in the control group. The three hypotheses are tested with this difference-in-difference (DID) methodology.

To observe knowledge transfer, we need to know the detailed content of knowledge and the information about senders and receivers. For this purpose, we use the patent application records submitted to the Japanese Patent Office. Although not every new piece of knowledge is documented as a patent application, they can be used as proxies (Acs, et al. (1994)). To trace knowledge transfer, we use patent citations. A pair of cited and citing patents enables us to know who are senders and receivers in the transfer, and when and where the transfer takes place from the bibliographic information of both patent applications. To classify these transfers into tow experimental groups, we retrieve micro-geographic information of locations of applicants, inventors, and the Shinkansen stations. In particular, the novelty of our data set is that patent application records are organized by their geographical proximities to the Shinkansen stations using locations occupied by all the applicants and inventors.

The paper is organized as follows. Section II describes the primary data sources used in the analysis including the Japanese patent database and geographic data sources. We also explain how to locate patent applications with respect to the Shinkansen stations and define notions to construct the treatment and control sample. Using the sample, Section III examines the three hypotheses of the study, and estimate the magnitude of the effect. Section IV discusses some issues to be further addressed, and concludes.

II. The Data

The basic data set is the IIP Patent Database (version 2015), which is the comprehensive database collecting the bibliographic information of patents applied for, and granted by, the Japan Patent Office from 1964 to 2012. From more than 10 million patent application records, we extract 9.9 million non-individual patent applications from 1976 to 2012 in which both applicants and inventors have addresses in Japan.

Because the patent bibliographic information of IIP DB is provided ‘as is’, we apply a series of text processing procedures to improve its quality. These include fixes to spelling errors found in the records and an identification of applicants and inventors. Our data to be used in the analysis must have information of where inventors were located when they applied for patents, which is unavailable in the IIP DB. We describe below how we construct our data.

The name field of the patent applicant data file is applied in an intensive cleanup process in which we generate a “standardized” notation of each applicant from its naming variations. This avoids us from a false negative problem where we mistakenly judge a pair of patents by the same applicant as by different ones due to their naming variations. We also give each applicant one of the following four types: individual, corporation, public institute, or academic institute, based on their names. Using these standardized applicant names, an applicant ID assigned by JPO, and NISTEP firm name dictionary (2014 version), we identified about 500,000 applicants in the period.

Geographical coordinates of applicants and inventors are obtained using their address fields of the inventor data file. Most of their addresses are specified at the block level, but it is written in free form. Here we also apply cleanup and standardization procedures to accommodate variations in notation, which gives us about 860,000 unique address. We use Microsoft BingMap service and “dams” by the Spatial Information Center, University of Tokyo[1] to assign a latitude and longitude to them. About 1.2% of those addresses are unidentified, but consequently for 9,668,647 patent applications both applicant and inventor addresses are assigned the coordinates successfully.

With the standardized notation of names and addresses, inventors are identified by the Computerized Matching Procedure (CMP, Trajtenberg, et al. (2006)). See Appendix A for details. The CMP identifies 2,148,476 domestic inventors who applied for patents between 1976 and 2012, and consequently, our patent data set has 20,826,111 inventor-patent pairs telling us which inventor applied for what patent at what time at which location in the period.

Patent citation records, by which we trace knowledge flows, are obtained from two data sources. The main source is the patent citation database “td5” provided by Tamada, et al. (2006)[2]. A notable feature of the database is that cited patents embedded in the body text of a citing patent are collected by a text scanning approach. These citations can be thought of as trails to the prior arts to which the applicants are “directly” referred, which we make use of. The secondary source is the citation database included in the IIP DB. The IIP DB provides three types of patent citation records: cited by applicants, examiners, or both. As the latter two citations are made by examiners to show an applicant the reasons to reject the application, we only use patent citations made by applicants, which are also considered as knowledge

[1] Our dams software are further modified to work with the location-reference information database by the National and Regional Policy Bureau of Japan (http://nlftp.mlit.go.jp/isj/) which gives geographic coordinates up to a block level.

[2] We use the td5 database licensed for the Institute of Innovation Research, Hitotsubashi University.

flows. Our patent citation data collected from these two sources consists of 9,721,089 records. We also notice that about 30% of them are self-citations in which cited and citing patents have at least one applicant or inventor in common. This type of citation does not represent knowledge transfer between different groups but refers to the past invention the citing applicant made. Therefore, as with many of the previous empirical studies on knowledge spillovers, self-citations are not included in the study.

II-1. Patent Locations and Definition of "On/Off the Line" Patents

The primary information in the sample construction is the location of a patent application, and its proximity to the Shinkansen stations. The information is based on the address information in patent documents and determined as follows. Each patent document includes the address information of applicants and inventors. While the applicant addresses are usually the headquarters of the applicant organization, the inventor addresses designate (in many cases) the places where the knowledge is born by the inventor, which can be different places than the applicant addresses. Therefore, we use inventor addresses to locate a patent application.

With this definition of patent location, we can know the geographical proximity between patent locations and the Shinkansen stations. To do so, for each inventor in a patent application, we first find the nearest station, one of the Shinkansen stations in operation (as of the patent application date) which the inventor reaches at the shortest time. We then calculate the road distance between the inventor's location and the nearest station[3]. Figure 1. shows the results from this "nearest station matching" procedure for the period 1998 to 2012 (after the opening of Hokuriku Shinkansen).

The figure shows that each inventor address is matched to one of the Shinkansen stations that allows the inventor to be reached in the shortest time. Frequency distributions of distances to the nearest station is shown in Figure 2.

We observe that, for all the Shinkansen lines, the first mode of the distribution is located at around 20 to 30 kilometers. This gives an idea as to the distance by which inventors are likely to use one of the Shinkansen lines to exploit knowledge in a distant area. Specifically, in this study, an inventor location is "near" one of the Shinkansen stations if its distance is less than or equal to the first mode of the distribution. To use this classification in a sample construction, we define the following two types for patent applications while using a threshold distance R (km). An "on-the-line" patent is defined as a patent in which all of the inventors are located within R kilometers from one of the Shinkansen stations. An "off-the-line" patent is defined as a patent in which none of the inventors are located within R kilometers from any of the Shinkansen stations. The threshold distance R is a device to split

[3] The traveling time to reach every Shinkansen station and the road distance (in kilometers) are calculated with the Open source routing machine (OSRM). The road route having the shortest time is found by Dijkstra's Algorithm.

Figure 1. Inventor locations and the nearest station (Period: 1998-2012)

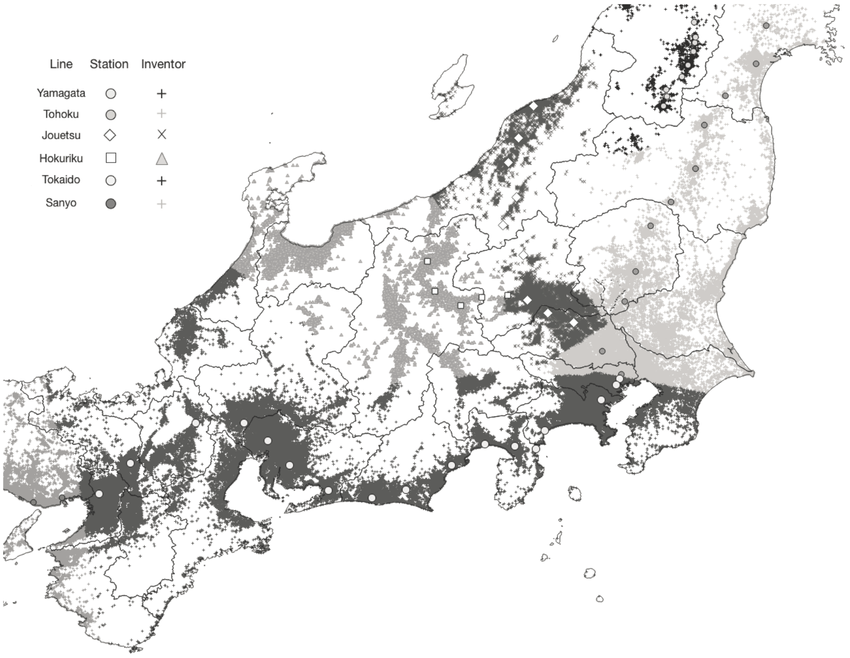


Figure 2. Frequency distributions of distances between an inventor location and its nearest station (by a Shinkansen line)

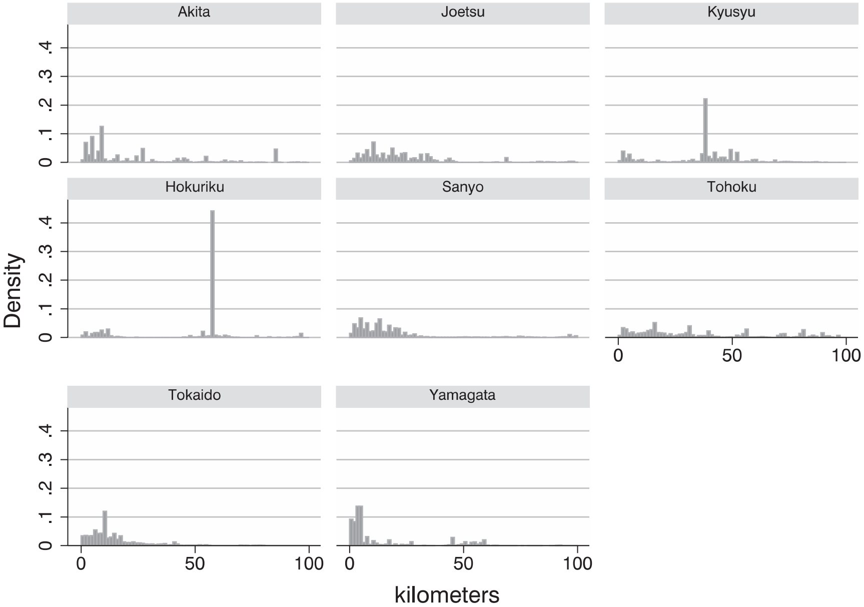
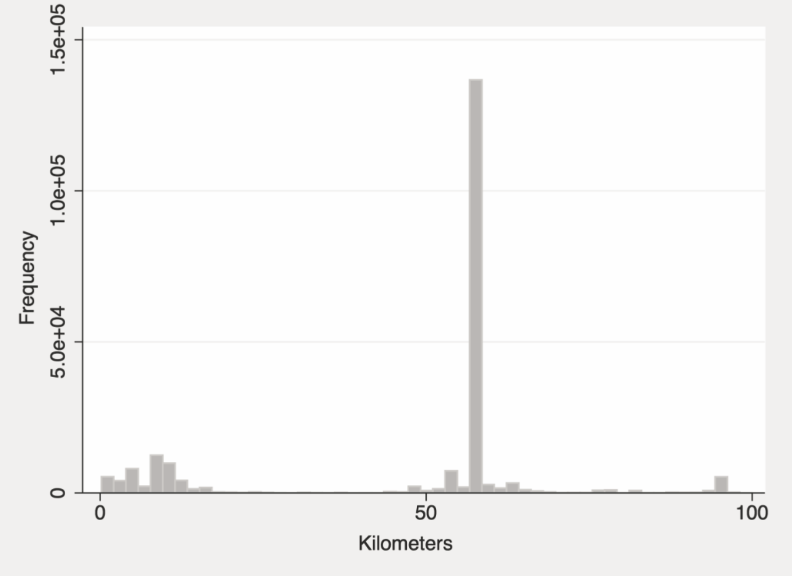


Figure 3. Frequency distributions of distances between an inventor location and its nearest station: the Hokuriku Shinkansen Line

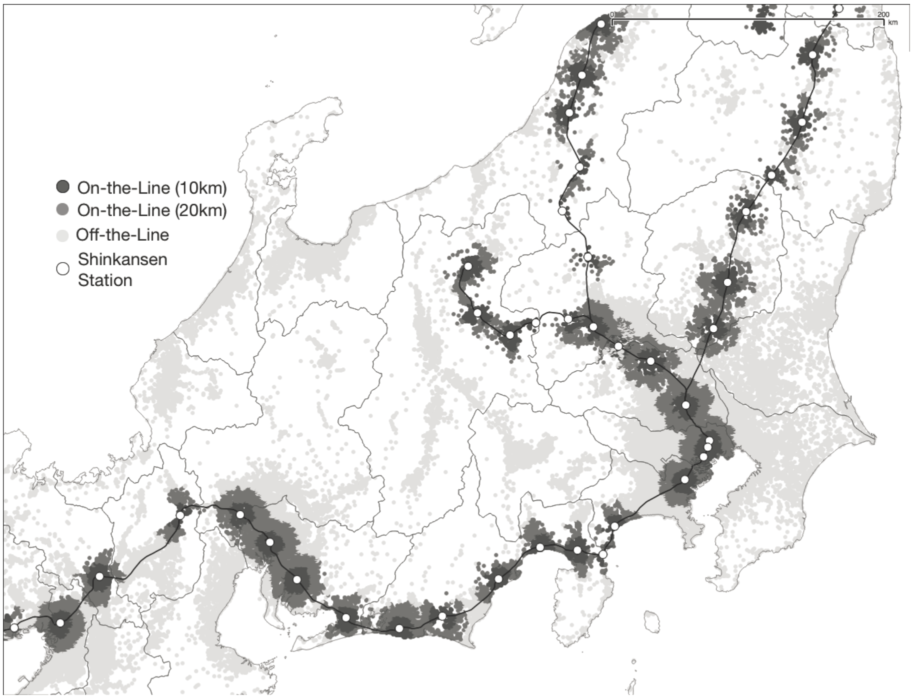


NOTE: The peak around 60km shows the patent applications from SEIKO-EPSON Corporation.

patent applications according to the proximity to the Shinkansen stations, which is given by a researcher. Therefore the choice of R affects the sample construction. Observing the distance distribution for inventors whose nearest stations is one of the Hokuriku Shinkansen stations (shown in Figure 3), we set R=10 kilometers.

Figure 4 shows all the inventor locations classified by R. Inventors at black (R=10km) and dark gray dots (R=20km) are located near the Shinkansen station. Inventors at gray dots are located at least 20km away from any of the Shinkansen stations. It may be seen that the choice of R=10km is too small and results in smaller sizes of the on-the-line sample. Also it is noted that patent applications do not meet the on/off requirement if some of the inventors are located near and the others are far. The smaller R is, the fewer the number of the on-the-line patents will be. However, as shown in Table 1 for R=10km, such cases are about 6.4% of the total patent applications.

Figure 4. A classification of inventor locations with respect to the Shinkansen Stations (period: 1998-2012).



NOTE: For each Shinkansen station (drawn as a white circle on the Shinkansen line), black (dark gray, resp.) dots indicate inventor locations distant from 10 (20, resp.) kilometers. The other dots (in light gray) are locations more than 20 kilometers away from any Shinkansen stations. The scale bar at the top right is 200 kilometers.

Table 1. The number of “on-the-line” and “off-the-line” patents for R=10km

Line Name	Start Year	End Year	Total	On	Off	(On+Off)/Total
Akita	1998	2012	4,779	2,154	2,141	89.87%
Joetsu	1983	2012	193,987	106,562	75,488	93.85%
Kyusyu	2005	2012	10,145	1,140	8,732	97.31%
Sanyo	1976	2012	790,509	399,128	326,605	91.81%
Nagano	1998	2012	220,787	30,660	182,768	96.67%
Tohoku	1983	2012	1,001,586	422,842	514,978	93.63%
Tokaido	1976	2012	7,431,086	5,541,755	1,421,684	93.71%
Yamagata	1998	2012	9,350	6,676	1,846	91.14%
All Lines	1976	2012	9,662,229	6,510,917	2,534,242	93.61%

II-2. Definition of "On/Off-the-line" Citations

Based on the definition of "on/off-the-line" patents, we define "on/off-the-line" citations, which tells us if the knowledge transfer as evidenced by a patent citation takes place on or off the Shinkansen line stations, respectively. An on-the-line citation is defined as a pair of cited and citing patents both of which are on-the-line patents. Similarly, an off-the-line citation is defined as a pair of cited citing patents both of which are off-the-line patents. The treatment and control group consists of on-the-line and off-the-line patents, respectively.

In order for the comparisons between those samples to capture only the effect of the opening of Shinkansen, the citation sample is further refined so that we only use citations made by inventors who have never moved between two samples throughout the period. Using the samples so constructed, we conduct a series of empirical analysis in the next section to see if the opening of the Hokuriku Shinkansen enhances innovative performance of inventors around the area. Table 2 shows the number of citations used in the following analysis.

Table 2. The Sample Size

	Pre-Period (Before the opening)	Post-Period (After the opening)
"On-the-line" Citations (Hokuriku Shinkansen)	565	4658
"Off-the-line" Citations	327658	603510

III. Empirical Analysis

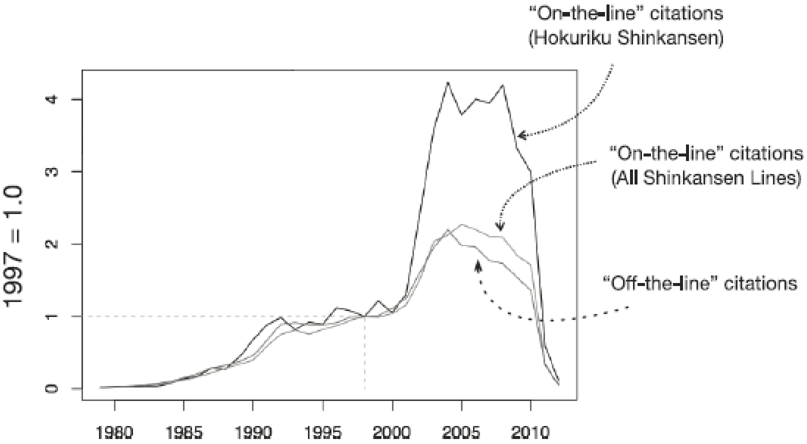
In this section, we examine the three hypotheses described in Section I. III-1 presents the result for the first hypothesis. For the period before and after the Hokuriku Shinkansen is built, we observe the changes in citation frequencies from the "on-the-line" *patents* for the Hokuriku Shinkansen line. To confirm those changes really occur in the frequencies of "on-the-line" *citations*, we group citations from the "on-the-line" patents into two subsamples by their destinations: if the opening of the Hokuriku Shinkansen causes the increase in citation frequencies, then we should observe that more citations are directed to "on-the-line" cited patents. That is, we should observe that "on-the-line" citations are more frequently observed than citations directed to "off-the-line" cited patents after the Shinkansen is built. This analysis reinforces the first hypothesis by showing the increase in the frequencies is indeed due to the opening of the Shinkansen.

In Section III-3, we apply difference in difference methodology to measure the effect the opening of the Hokuriku Shinkansen on the geographical expansion of "on-the-line" citations from the "on-the-line" patents for the Shinkansen. The first half of III-3 tests the second hypothesis by estimating the average treatment effect for the expansion of citation distance, and the second half examines the third hypothesis by measuring the quantile treatment effect. We obtain the estimates at each quantile before and after the opening of the Hokuriku Shinkansen, and observe which part of the changes can contribute the changes in the average treatment effect.

III-1. Citation Frequency Results

Figure 5 presents relative citation frequencies of "on-the-line" citations for the Hokuriku Shinkansen lines, "on-the-line" citations for all the Shinkansen lines and "off-the-line" citations (from an "off-the-line" patent to an "off-the-line" patent), scaled by the values in 1997. All lines in the figure exhibit an increasing trend, but as is clearly shown in the figure, the frequencies of "on-the-line" citations for the Hokuriku Shinkansen lines jump up after a few years have passed since the opening of the Hokuriku Shinkansen. The lag of about 4 years between the opening and the big increase in the figure may reflect the fact that new research projects generally take a while to have some results. Therefore the increase of "on-the-line" citations for the Hokuriku Shinkansen may be due to its opening.

Figure 5. Time series plots of citation frequencies

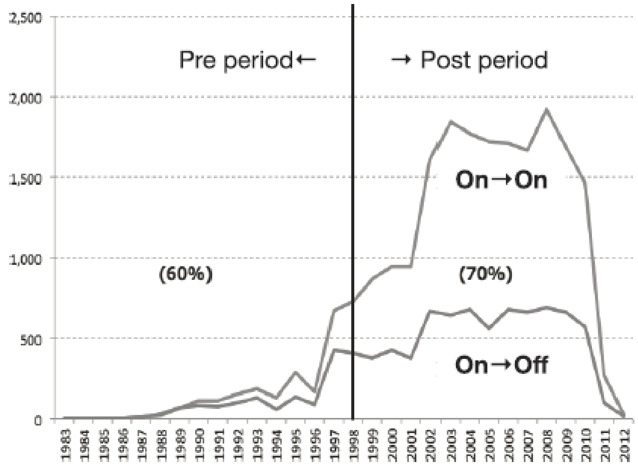


NOTE: Each line is scaled by the 1997 value

III-2. Direction of "on-the-line" Patents

Figure 6 shows time series plot citations from "on-the-line" patents, each line representing citing frequencies to cited "on-the-line" patents and cited "off-the-line" patents. Again citation to cited "on-the-line" patents much more increases after the opening of the Hokuriku Shinkansen than the other, suggesting that the opening has significant effect on the increasing

Figure 6. Directions of “on-the-line” citations



NOTE: On→On represents citations between “on-the-line” patents. On→Off represents citations from “on-the-line” patents to “off-the-line” patents.

in “on-the-line” citation frequencies. The fraction of “on-the-line” citations increases from 60% to 70% after the opening of the Shinkansen.

III-3. Citation Distance Results

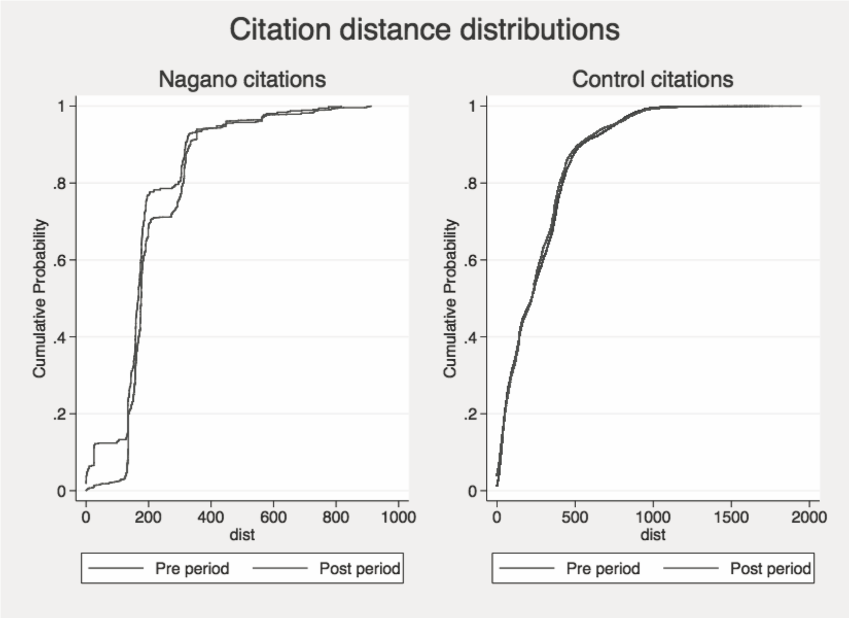
Now we turn to test hypothesis 2 and 3 using difference-in-difference (DID) methodology. If the opening of the Hokuriku Shinkansen does not have any effect on the geographical extent of knowledge transfer, we should observe that the distances of “on-the-line” citations before and after the opening of Hokuriku Shinkansen do not have significant differences, with controlling for other factors affecting the distances. To investigate this, we take “off-the-line” citations as the control group and compare their differences in distances with those of the “on-the-line” citations (the treatment group), before and after the opening. The DID results are shown in Table 3. The estimate, the average treatment effect of the opening, is positive (35km) and significant (t-value is 7). Therefore the hypothesis is positively supported. We conclude that the opening of the Hokuriku Shinkansen expands the geographical extent of knowledge transfer by 35km, on average.

To examine the last hypothesis, we apply quantile DID (QDID) to the changes in citation distances before and after the opening of Hokuriku Shinkansen. In addition to the DID results obtained above, QDID further tells us which part of citation distance distribution has significant change before and after the opening. To begin with, we plot in Figure 7 the citations distance distributions of treatment and control sample before and after the opening of the Shinkansen.

Table 3. Difference-in-difference Estimation Results

	Before the opening			After the opening			
	"Off-the-line"	"On-the-line"	Difference	"Off-the-line"	"On-the-line"	Difference	Difference in Difference
Citation Distance	261.61	191.11	-70.51	255.43	219.97	-35.46	35.04
Standard Error	0.39	5.42	5.44	0.29	1.81	1.84	5.74

Figure 7. Distribution of citation distances before and after the opening of the Hokuriku Shinkansen

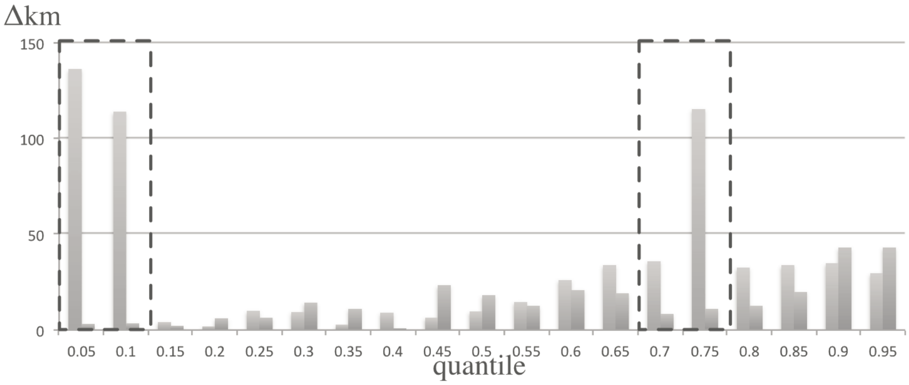


NOTE: The treatment group: left, the control group: right

"Off-the-line" citation distance distribution remains unchanged, but the "on-the-line" citation distance distribution changes its shape so that within 100km citation distances are fewer in the post period (after the opening). To confirm these observations, we conduct QDID and the result is shown in Figure 8.

The estimation results show that the distribution shifts by 100 to 150km at 5% and 10% percentiles. Also the distribution shifts by 40km at 70%, and 120km at 75% percentiles. The shift at low distances clearly indicates that the shorter citation distances become relatively few in the post period. If we map this change in the geography around the Hokuriku Shinkansen stations, it is observed that the direction of the citation changes from Nagano area to the area around Tokaido Shinkansen stations (Shin-Yokohama station, in particular). Hence, it can be concluded that the opening of the Hokuriku Shinkansen enable knowledge transfer received in the area to shift from within the area to outside the area. Another shift at 70% and 75% percentiles is made by citations to "on-the-line" patents around the Tokaido,

Figure 8. Quantile difference-in-difference results



NOTE: At each quantile (from 0.05 to 0.95), the height of left and right bar is the estimate and the standard error, respectively.

Sanyo, and Tohoku Shinkansen lines, which should be explored further by examining a sender and receiver of each citation.

IV. Concluding Remarks

This paper studies the effect of high-speed railways on knowledge transfer as evidenced by patent citations. We design a natural experiment that uses the opening of the Hokuriku Shinkansen in 1997 as the exogenous event. The treatment group is the set of patent citations from inventors located within 10km from the Hokuriku Shinkansen stations. The control group is the set of patent citations from inventors whose locations are more than 10km away from any Shinkansen station. Using the difference-in-difference methodology, we find that the opening of the Hokuriku Shinkansen increases the frequency of knowledge transfer, expands the geographical extent of knowledge transfer by 35km, and promotes knowledge transfer from further areas where other Shinkansen lines run. These empirical results suggest that the high-speed railway service, a form of social capital, helps to enhance the innovative productivity in a region distant from the metropolitan area, by reducing the cost to access new knowledge.

We exclude citations made by firms or inventors that move between the treatment and control groups during the period. This is because we keep one of the basic assumptions with which a natural experiment works. Therefore, we exclude alternative channels of knowledge transfer: Firms or inventors can move into the area that they expect to enhance their innovative productivity by the opening of high-speed railways and knowledge transfer takes place by these moves. The study should develop a framework to account for this type of knowledge transfer.

References

Acs, Zoltan J., David B. Audretsch, and Maryann P. Feldman (1994) "R&D spillovers and recipient firm size," *The Review of Economics and Statistics*, vol. 76, no. 2, pp. 336-340.

Acs, Zoltan, Luc Anselin, and Attila Varga (2002) "Patents and innovation counts as measures of regional production of new knowledge," *Research Policy*, vol. 31, pp. 1069-85.

Agrawal, Ajay, Cockburn, Iain, McHale, John (2006), "Gone but not forgotten: Labor flows, knowledge spillovers, and enduring social capital," *Journal of Economic Geography*, vol. 6, no. 5, pp 571-591.

Agrawal, Ajay, Devesh Kapur, and John McHale (2008), "How do spatial and social proximity influence knowledge flows? Evidence from patent data," *Journal of Urban Economics*, vol. 64, pp. 258-69.

Arrow, Kenneth, J (1962) "Economic welfare and the allocation of resources for invention," in *The rate and direction of inventive activity*, ed. by R. Nelson, pp. 609- 626. Princeton University Press, Princeton, NJ.

Goto, Akira, and Kazuyuki Motohashi (2005) "Tokkyo Database no Kaihatsu to Innovation Kenkyuu," Chizaiken Forum, vol. 63, p. 43. (trans: Development of Patent Database and the Study on Innovation)

Griliches, Zvi (1979) "Issues in Assessing the Contribution of R&D to Productivity Growth," *Bell Journal of Economics,* vol. 10, pp. 92-116.

Jaffe, Adam B (1986) "Technological Opportunity and spillovers of R&D," *American Economic Review*, vol. 76, pp. 984-1001.

Jaffe, Adam B (1989) "Real Effects of Academic Research," *American Economic Review*, vol. 79, no. 5, pp. 957-70.

Jaffe, A.B., M. Trajtenberg, and R. Henderson (1993) "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations," *The Quarterly Journal of Economics*, vol. 108, no. 3, pp. 577-598.

Murata, Yasusada, Ryo Nakajima, Ryosuke Okamoto, and Ryuichi Tamura (2014), "Localized knowledge spillovers and patent citations: A distance-based approach", *Review of Economics and Statistics*, vol. 96, pp. 967-985.

Nakajima, Ryo, Ryuichi Tamura, and Nobuyuki Hanaki (2010) "The effect of collaboration network on inventors' job match, productivity and tenure." *Labour Economics*, vol. 17, pp. 723-734.

Nakamura, Kenta, *IIP patent database manual*, 2015/7.

Saito, Yukiko Umeno, and Isamu Yamauchi (2015) "Inventors' Mobility and Organizations' Productivity: Evidence from Japanese rare name inventors," *RIETI Discussion Paper Series 15-E-128*.

Saxenian, Anna Lee, *Silicon Valley's new immigrant entrepreneurs*, Public Policy Institute of California, 1999.

Schumpeter Tamada, Yusuke Naito, Kiminori Gemba, Fumio Kodama, and Jun Suzuki

(2006), “Significant Difference of Dependence upon Scientific Knowledge among Different Technologies,” *Scientometrics*, Vol. 68, No. 2, pp. 289-324.

Singh, Jasjit (2005) “Collaborative networks as determinants of knowledge di useion patterns,” *Management Science*, vol. 51, no. 5, pp. 756-770.

Stefano Breschi, and Francesco Lissoni (2009) “Mobility of skilled workers and co-invention networks: an anatomy of localized knowledge flows,” *Journal of Economic Geography*, vol. 9, pp. 439-468.

Thompson, Peter, and Melanie Fox-Kean (2005). “Patent citations and the geography of knowledge spillovers: A reassessment.” *American Economic Review*, 95: 450-460.

Trajtenberg, Manuel, Gil Shiff, and Ran Melamed (2006) “The NAMES GAME: Harnessing inventors’ patent data for economic research.” *NBER Working Paper #12479*.

V. Appendix

V-1. Name Matching Algorithm for Japanese Inventor Names

The analysis in the study heavily relies on the identification information of applicants and inventors in the Japanese patent bibliography. For the former, there have been a couple of data sources available to us (Applicant number ID given by the Japanese Patent Office (JPO), and NISTEP applicant name dictionary). Both help us to identify more than 90% of Japanese applicant organizations found in applicant fields of the patent documents. But for the latter, JPO does not issue any identifier for an inventor[4]. Therefore it is our responsibility to develop a systematic identification assignment scheme using Japanese patent bibliographic information. Since the total number of inventors in Japanese patent applications is more than 10 million, it is impractical to identify each inventor manually. A series of name-matching procedures using computers has thus been developed which automatically identifies an inventor by comparing pairs of patent documents. In their early stage of development, the procedures try to identify an inventor by mere coincidence of her family and given names found in the inventor name fields between patent documents, but such a “naïve” name-matching method could introduce what is called a “false positive” problem in which a researcher mistakenly judges different inventors having the same name as the same inventor. The Computerized Matching Procedure (CMP) by Trajtenberg, et al. (2006) allows us to get around this problem by splitting a set of patent applications of inventors having the same names into subsets of applications according to matches of the other patent bibliographic information, each of which is then regarded as a collection of applications by different

[4] The IIP patent database as of 2014 had provided an identification number for each inventor (“row” field in the database) derived by a unique combination of a family name, a given name and an inventor address. However, as stated in the following, those field values are not standardized and thus variations in notation in the fields could lead to a “false negative” problem where two applications from the same inventor are judged as those from different inventors simply due to the notational variations.

inventors. This study uses a variant of the CMP for Japanese patent data and identifies about 800,000 inventors living in Japan. The CMP was originally developed for identifying US inventors and used in several empirical studies, an implementation of which is detailed in Nakajima, et al. (2010). In this appendix, we outline a variant of the CMP which is modified to work for Japanese patent documents written in the Japanese language. Our focus is twofold. First, how our procedure is localized in a Japanese language environment. Second, how our procedure deals with a "triangular violation" issue that can arise in pair-wise matching between different patent documents.

V-2. Outline of Computerized Matching Procedure for Japanese inventors

The Computerized Matching Procedure (CMP) is a systematic way to identify patent inventors using not only name information, but also other patent attributes in the application that can be helpful for the identification. The procedure consists of the following two efforts, each of which aims to control for one of two mistakes which a researcher can make for the identification:

(A) To reduce a possibility that the same inventor having several patent applications is judged as different inventors simply due to errors or notational variations in each bibliography ("false negative" problem), patent bibliographic information including names, address are converted to their standardized notation.

(B) To reduce a possibility that different inventors having the same name are judged as the same inventor ("false positive" problem), patent applications having the same name are compared using other patent bibliographic information such as patent technology classification, co-authorship, and citations to previous patent application.

To implement (A) and (B), our procedure is organized with the following two steps:

[STEP 1]

The procedure begins with cleanup work for the bibliographic information in patent applications. In this work, consecutive white spaces, phonetic characters, and numeric notations in Kanji or Arabic character are converted to single notation. An address notation is converted to its up-to-date address as of 2014. These standardization reduces notational variations for a single noun, and thus reduces the "false negative" errors. Using standardized names, patent applications are grouped into collections of patent applications that have the same family/given names. Following Trajtenberg, et al. (2006) we call each collection *P-Set*. For about 12 million of inventor-patent pairs applied in 1976-2013, we construct about 1.65 million P-Sets.

[STEP 2]

Each P-Set consists of patent applications by inventors with the same names. In Step 2, we split the P-Set by pairwise matching of applications and a scoring method. Specifically, we first set a threshold value, and if the score exceeds this value, the two applications belong to the same inventor. The scoring system starts from 0 points when the pair match begins. We then compare address, patent technology class, backward citations, and coauthor attributes to each other, and add scores if they match. If the total score is below the threshold value, the pair is judged to belong to different inventors (Failure), otherwise the pair is judged to belong to the same inventor (Success). Figure A-1 shows the match result indicating successes (1) and failures (0). Repeating this matching process for every P-Set, we obtain about 2.17 million unique inventors.

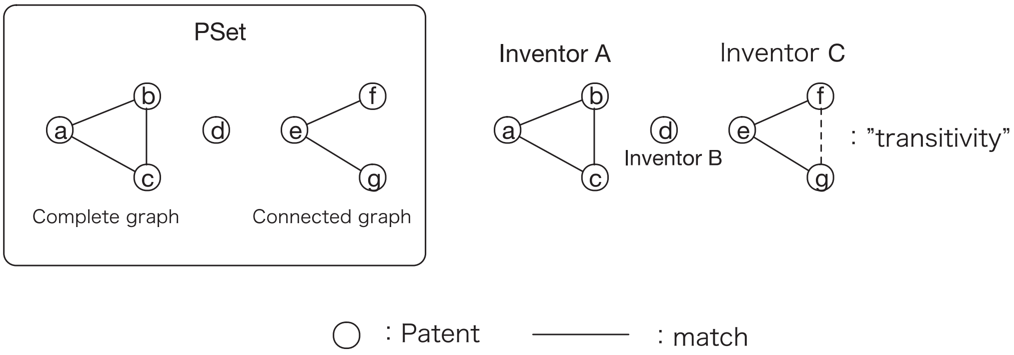
Figure A-1. Match results for a P-Set of 7 applications

	a	b	c	d	e	f	g
a							
b	1						
c	1	1					
d	0	0	0				
e	0	0	0	0			
f	0	0	0	0	1		
g	0	0	0	0	1	0	

V-3. Force Match via Transitivity

In practice, we frequently encounter the case where a pairwise match produces "inconsistent" results. In the match results in Figure A-1, the result for applications a, b, and c is consistent: a matches b, b matches c, and c matches a. However, the result for applications e, f, and g is not consistent: e matches f, e matches g, but f does not match with g. This is a famous issue known as a "triangular violation", and following Trajtenberg, et al. (2006) we force a match between f and g by transitivity. In implementation, finding a match via transitivity is equivalent to finding a connected graph in the match results. This is easily handled by computer libraries supporting the graph theory. We used the igraph library (http://igraph.sf.net).

Figure A-2. Matching via Transitivity


PSet
a
b
c
d
e
f
g
Complete graph
Connected graph
Inventor A
Inventor B
Inventor C
: "transitivity"
: Patent
: match